NASA CR-201715

THE PSEUDO-INVERSE OF THE DERIVATIVE OPERATOR IN POLYNOMIAL SPECTRAL METHODS

Josip Loncaric

The matrix $D-kI$ in polynomial approximations of order N is sin

to a large Jordan block which is invertible for nonzero k but extremely sensitive to perturbation. Solving the problem $(D-kI)f = g$ involves similarity transforms whose condition number grows as $N!$, which exceeds typical machine precision for $N > 17$. By using orthogonal projections, we reformulate the problem in terms of Q, the pseudo-inverse of D, and therefore its optimal preconditioner. The matrix Q in commonly used Chebyshev or Legendre representations is a simple tridiagonal matrix and its eigenvalues are small and imaginary. The particular solution of $(I - kQ)f = Qg$ can be found for all real k at high resolutions and low computational cost $(O(N)$ times faster than the commonly used Lanczos tau method). Boundary conditions are applied later by adding a multiple of the known homogeneous solution. In Chebyshev representation, machine precision results are achieved at modest resolution requirements. Multidimensional and higher order differential operators can also take advantage of the simple form of Q by factoring or preconditioning.

17 pages
July 1997

THE PSEUDO-INVERSE OF THE DERIVATIVE OPERATOR IN POLYNOMIAL SPECTRAL METHODS

Josip Lončarić*
ICASE
NASA Langley Research Center
Hampton, VA 23681-0001

Abstract

The matrix $D - kI$ in polynomial approximations of order N is similar to a large Jordan block which is invertible for nonzero k but extremely sensitive to perturbation. Solving the problem $(D - kI)f = g$ involves similarity transforms whose condition number grows as $N!$, which exceeds typical machine precision for $N > 17$. By using orthogonal projections, we reformulate the problem in terms of Q, the pseudo-inverse of D, and therefore its optimal preconditioner. The matrix Q in commonly used Chebyshev or Legendre representations is a simple tridiagonal matrix and its eigenvalues are small and imaginary. The particular solution of $(I - kQ)f = Qg$ can be found for all real k at high resolutions and low computational cost ($O(N)$ times faster than the commonly used Lanczos tau method). Boundary conditions are applied later by adding a multiple of the known homogeneous solution. In Chebyshev representation, machine precision results are achieved at modest resolution requirements. Multidimensional and higher order differential operators can also take advantage of the simple form of Q by factoring or preconditioning.

*This research was partially supported by the National Aeronautics and Space Administration under NASA Contract No. NAS1-19480 while the author was in residence at the Institute for Computer Applications in Science and Engineering(ICASE), M/S 403, NASA Langley Research Center, Hampton, VA, 23681-0001.

1 Introduction

In many situations it is necessary to solve a family of differential equations of the form $\mathcal{A}_k f = g$ where $\mathcal{A}_k$ is a differential operator, k is a parameter, and $f \in D(\mathcal{A}_k)$ and $g \in R(\mathcal{A}_k)$ are functions of the independent variable $x \in \Omega$. A common example of this problem is the differential equation

$$\left(\frac{d}{dx} - k\right) f(x) = g(x) \tag{1}$$

where k is real and the absolute value of k is bounded by a possibly large number K. The domain on which the differential equation holds is a finite interval, say $x \in (-1, 1)$, and the Dirichlet boundary condition $f(x_b) = 0$ is applied at the boundary $x_b \in \{\pm 1\}$. Since the problem admits the homogeneous solution $f_h(x) = e^{k(x-x_b)}$ which can rapidly grow in the $\mathrm{sgn}(k)$ direction, it is essential to apply the boundary condition at the proper side of the domain ($x_b = \mathrm{sgn}(k)$) in order to have a well posed problem.

Our choice of the model problem (1) is motivated by physical applications, but also by the fact that the polynomial approximation of (1) has the form of a Jordan block matrix. Since any matrix can be written as $A = S + N$ where S is diagonalizable, N is nilpotent and $SN = NS$, the correspondence between (1) and its polynomial approximation is of fundamental importance. The well known numerical difficulties with Jordan blocks of a matrix are reflected in solving the problem (1).

We shall show that continuous and discrete versions of the problem (1) may be replaced by a reformulated optimally preconditioned problem whose solution involves operations with simple tridiagonal matrices. The reformulated problem leads to a solution procedure which is very efficient, numerically stable, and accurate to machine precision ($\epsilon \approx 2.22 \times 10^{-16}$ when double precision is used) in Chebyshev polynomial representation at modest resolution requirements which grow slowly with $|k|$.

1.1 An example

In cylindrical geometry, the Laplacian operator may be written as

$$\nabla^2 = \frac{1}{r^2}\left[\left(\frac{\partial}{\partial \log(r)}\right)^2 + \left(\frac{\partial}{\partial \theta}\right)^2\right]. \tag{2}$$

Taking the Fourier transform in the θ direction, we may write

$$\begin{aligned}
\nabla_k^2 &= \frac{1}{r^2}\left[\left(\frac{d}{d\log(r)}\right)^2 - k^2\right] \\
&= \frac{1}{r^2}\left[\frac{d}{d\log(r)} - |k|\right]\left[\frac{d}{d\log(r)} + |k|\right] \\
&\stackrel{\text{def}}{=} \frac{1}{r^2}\mathcal{B}\mathcal{F}
\end{aligned} \tag{3}$$

where k is the circumferential wavenumber and $\mathcal{F}$ and $\mathcal{B}$ are operators of the form (1) with $x = \log(r)$. This factorization of the Laplacian arises in the context of applying the exact

artificial boundary conditions [6]. To prohibit the growing mode as $r \to \infty$, one must have that

$$\left[\frac{d}{d\log(r)} + |k|\right] f_k(r)\Big|_{r=R} = 0 \tag{4}$$

at the limit $r = R$ of the finite computational domain. This boundary condition is nonlocal and it is best applied in the Fourier domain. To solve the problem

$$\nabla^2 \psi = -\omega \tag{5}$$

with the boundary conditions $\psi|_{r=1} = 0$ and $\psi \to 0$ as $r \to \infty$, assume that the support of ω is contained within the computational domain of radius R. Given that $\omega = 0$ outside the computational domain, the boundary condition (4) applied at $r = R$ is the exact equivalent of the boundary condition at infinity. The solution in the Fourier domain takes three steps and may be written as

$$\zeta_k = -r^2 \omega_k \tag{6}$$

$$\xi_k = \mathcal{B}^{-1} \zeta_k \text{ where } \xi_k(r)|_{r=R} = 0 \tag{7}$$

$$\psi_k = \mathcal{F}^{-1} \xi_k \text{ where } \psi_k(r)|_{r=1} = 0 \tag{8}$$

where the inverses of the operators $\mathcal{F}$ and $\mathcal{B}$ are taken by applying the indicated boundary conditions. Therefore, this numerical solution depends on solving the equation (1).

2 Polynomial approximation

Our aim shall be to solve this equation numerically on function spaces consisting of polynomials of degree at most N. Let D be the discretized representation of the derivative operator on the $(N+1)$-dimensional space P_N of N-th order polynomials, and let us write the original equation as follows:

$$(D - kI)f = g \tag{9}$$

where f and g are vectors representing the functions $f(x)$ and $g(x)$ in P_N. The boundary condition also has to be satisfied, but as we shall see, this leads to an overdetermined problem whenever $k \neq 0$. In this regard, polynomial approximations are intrinsically different from the original continuous problem.

There are two difficulties that have to be resolved. First, each differentiation reduces the degree of the polynomial, cascading all polynomials in P_N to zero in at most $N + 1$ differentiation steps. Therefore, D is a nilpotent operator with the characteristic polynomial $\chi(s) = s^{N+1}$, which is also its minimal polynomial. It follows that for all $k \neq 0$ the matrix $D - kI$ is invertible and our representation has a unique solution, which is in general incompatible with the boundary condition.

Since $\chi(D) = D^{N+1} = 0$, we can write this solution exactly as a finite sum when $k \neq 0$:

$$f = -\frac{1}{k} \sum_{n=0}^{N} \left(\frac{D}{k}\right)^n g \tag{10}$$

This equation, while exact, is extremely poorly conditioned for large N; this constitutes the second difficulty [5]. By changing the polynomial basis functions from x^n to $x^n/n!$, we see that the above formula defines an upper triangular Toeplitz matrix T with entries $T_{ij} = -k^{i-j-1}$ for $j \geq i$. However, the similarity transform between the basis functions x^n and $x^n/n!$ has the condition number $N!$, which appears singular to machine precision when $N > 17$. Therefore, this exact equation is not computationally useful in the context of solving differential equations where N needs to be much larger than just 17.

This problem becomes even harder when orthogonal polynomials are used. The condition number of the similarity transform between the Chebyshev polynomial basis $\{T_n(x)\}_{n=0}^N$ and the companion form basis $\{x^n/n!\}_{n=0}^N$ is $2^{N-1}N!$, which exceeds machine precision as soon as $N > 14$.

Both difficulties can be traced to the loss of the homogeneous solution in the discretized problem. More degrees of freedom are needed to resolve the difficulties. The Lanczos tau method [4] introduces an additional high order polynomial of degree $N+1$ in order to satisfy the boundary condition. The resulting system of equations on P_{N+1} may be written as $(A + uv^T)f = (g^T, 0)^T$, where A restricted to P_N coincides with $D - kI$ and the outer product uv^T arises from the boundary condition. Even though the matrix A may have a simple structure, $A + uv^T$ is hard to invert since one cannot directly apply the Sherman-Morrison formula [3]

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u} \tag{11}$$

due to the fact that inverting the restriction of A to P_N is ill-conditioned.

We propose a procedure which is well conditioned and still preserves the tridiagonal form of the matrices involved. This approach is $O(N)$ times faster than the the tau method which introduces a full row into the matrix to be inverted.

3 Problem reformulation

The source of difficulties is the nilpotent part of $D - kI$. We note that D is a non-normal operator for which the commutator $DD^* - D^*D$ is of rank 2 in companion form representation and acts by

$$DD^* - D^*D : \sum_{n=0}^{N} \frac{a_n x^n}{n!} \longmapsto a_0 - \frac{a_N x^N}{N!} \tag{12}$$

so it is natural to use a preconditioner based on D. Fortunately, the optimal preconditioner in typical spectral representations has a simple form. This important observation suggests the following solution method.

Instead of using the Lanczos tau method to apply the boundary condition, we shall construct an integration operator Q as the pseudo-inverse of D and seek a particular solution of the discretized problem rewritten as follows:

$$(I - kQ)f_p \approx Qg = Q(D - kI)f. \tag{13}$$

The approximation $QD \approx I$ will be justified later, in section 6. The reformulated problem $(I - kQ)f_p = Qg$ has the solution $f_p = (I - kQ)^{-1}Qg$. The full solution of (9) is

then obtained as $f = f_p + \alpha f_h$ where f_h represents the homogeneous solution $e^{k(x-x_b)}$ and $\alpha = -f_p(x_b)$. This delayed application of the boundary condition allows us to choose an integration operator Q with optimal numerical properties, which can be thought of as a surrogate boundary condition. This two step solution procedure is analogous to the use of the Sherman-Morrison formula. The reformulated problem now involves integrations instead of differentiations. This change leads to well behaved numerical implementations. Moreover, we shall demonstrate that in typical polynomial representations the matrix Q has a simple form which leads to efficient algorithms.

Integration preconditioners for differential operators in spectral methods are tridiagonal for arbitrary classical orthogonal polynomial families [2]. The use of the pseudoinverse of D, proposed here, is an improvement which simplifies analysis and guarantees optimal numerical results in each orthogonal polynomial representation.

4 The pseudo-inverse of the derivative operator

The pseudo-inverse Q of the matrix D is the unique minimal Frobenius norm solution to $\min \|DQ - I\|_F$. This condition amounts to the requirement that DQ and QD be orthogonal projections onto image(D) and image(Q), respectively [3].

To construct the pseudo-inverse of D we first note that $\ker(D) = P_0$ and that image(D) = $P_{N-1} \subset P_N$. From the commutative diagram

$$\begin{array}{ccc} P_N & \xrightarrow{D} & P_N \\ \Pi \downarrow & & \uparrow \Phi \\ P_N/P_0 & \xrightarrow{\Delta} & P_{N-1} \end{array} \tag{14}$$

where Π is the canonical projection to the coset space P_N/P_0, Δ is invertible, and Φ is the insertion map, we conclude that the pseudo-inverse $D^{(-1)}$ is given by

$$Q \stackrel{\text{def}}{=} D^{(-1)} = \Pi^{(-1)} \Delta^{-1} \Phi^{(-1)} \tag{15}$$

where $\Phi^{(-1)}$ is the orthogonal projection to P_{N-1} and $\Pi^{(-1)}$ is the insertion map whose image is $P_0^{\perp}$, the subspace orthogonal to P_0. *In this construction orthogonality plays a key role.* Each Q depends on a chosen inner product on P_N, which will be clear from the context.

This definition satisfies the conditions that QD and DQ be the orthogonal projections to image(Q) = $P_0^{\perp}$ and image(D) = P_{N-1}, respectively. The commutator is

$$[D, Q] = DQ - QD = \Pi_{P_{N-1}} - \Pi_{P_0^{\perp}} = \Pi_{P_0} - \Pi_{P_{N-1}^{\perp}} \tag{16}$$

which is an operator of rank 2, showing that Q and D nearly commute. Orthogonal projection operators Π_V to various subspaces V will be appear often in the following discussion.

Let $\{p_n\}_{n=0}^N$ be orthogonal basis polynomials (not necessarily normalized) such that $\deg(p_n) = n$. The matrix form of D is

$$D = \begin{bmatrix} 0_{N\times 1} & \Delta \\ 0_{1\times 1} & 0_{1\times N} \end{bmatrix} \tag{17}$$

and therefore

$$Q = \begin{bmatrix} 0_{1\times N} & 0_{1\times 1} \\ \Delta^{-1} & 0_{N\times 1} \end{bmatrix} \tag{18}$$

This gives us a simple method of constructing Q. For example, if the polynomials $p_n(x) = x^n/n!$ are defined to be orthogonal, D is reduced to its companion form where $\Delta = I$ and $Q = D^T$.

For the metric generated by the Chebyshev polynomials $p_n(x) = T_n(x) = \cos(n\cos^{-1}(x))$, one obtains the tridiagonal matrix

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 & & \cdots & & 0 \\ 1 & 0 & -\frac{1}{2} & 0 & & & & \\ 0 & \frac{1}{4} & 0 & -\frac{1}{4} & & & & \\ & \ddots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & & \\ & & & \frac{1}{2(N-2)} & 0 & -\frac{1}{2(N-2)} & 0 & \\ & & & 0 & \frac{1}{2(N-1)} & 0 & 0 & \\ 0 & \cdots & & 0 & 0 & \frac{1}{2N} & 0 & \end{bmatrix} \tag{19}$$

while the Legendre polynomials lead to a slightly different tridiagonal matrix

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 & & \cdots & & 0 \\ 1 & 0 & -\frac{1}{5} & 0 & & & & \\ 0 & \frac{1}{3} & 0 & -\frac{1}{7} & & & & \\ & \ddots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & & \\ & & & \frac{1}{2N-5} & 0 & -\frac{1}{2N-1} & 0 & \\ & & & 0 & \frac{1}{2N-3} & 0 & 0 & \\ 0 & \cdots & & 0 & 0 & \frac{1}{2N-1} & 0 & \end{bmatrix} \tag{20}$$

Numerical evidence indicates that the eigenvalues of Q in both Chebyshev and Legendre representations lie on the imaginary axis. This favorable distribution of eigenvalues need not happen in general. We conclude that in all three cases (companion form, Chebyshev and Legendre representations) the matrix $I - kQ$ is an invertible tridiagonal system for all real k.

5 Eigenvectors and eigenvalues of Q

Given an orthonormal basis of P_N ordered by basis polynomial degree, let us denote the components of a vector $v \in P_N$ by $v_0, v_1, \ldots, v_N$. In any polynomial representation, for eigenvalue $\lambda \neq 0$ the equation $Qv = \lambda v$ implies that the constant polynomial component satisfies $v_0 = 0$ so that $v \perp P_0$. Thus, we obtain

$$DQv = \lambda Dv \tag{21}$$

where DQ is an orthogonal projection to P_{N-1}, the range of D. Recall that $p_N \perp P_{N-1}$ and scale v so that $v = u + p_N$ where $u \in P_{N-1}$. Therefore, $DQv = v - p_N$ and

$$(I - \lambda D)v = p_N \tag{22}$$

so that

$$v = (I - \lambda D)^{-1} p_N = \sum_{n=0}^{N} \lambda^n D^n p_N \tag{23}$$

where $v \perp P_0$ provided that λ is an eigenvalue of Q. This exact equation defines the eigenvectors once the nonzero eigenvalues of Q have been obtained, but it is very sensitive to perturbations in λ due to the same numerical difficulties as the equation (9). Conversely, eigenvalues λ are very sharply defined. Our aim shall be to determine the form of eigenvectors analytically.

We note that the homogeneous solution of the continuous equivalent of the eigenvector equation (22) is $e^{x/\lambda}$, a highly oscillatory function for small but nonzero imaginary λ. One is reminded of the Fourier series, since the frequencies $|1/\lambda|$ are approximately evenly spaced given Chebyshev representation. The integration operator Q in Chebyshev representation may be thought of as a rough approximation of the Fourier integration operator. This suggests that the Jordan decomposition of Q may be reasonably well conditioned.

A particularly simple form of v can be derived when $N + 1 = 2^J$. In this case, the sum of $N + 1$ terms can be factored as a product of only $\log_2(N + 1)$ terms so that

$$v = \prod_{j=0}^{J-1} (I + (\lambda D)^{2^j}) p_N. \tag{24}$$

The eigenvalue zero has multiplicity at least one, since $Qp_N = 0$. A basis for the invariant subspace associated with $\lambda = 0$ can be constructed by seeking nontrivial solutions to equations of the form $Q^j v_j = 0$. For basis polynomials $x^n/n!$ which bring D to its companion form, all eigenvalues of Q are zero, and Q acts by mapping $x^n/n! \mapsto x^{n+1}/(n+1)!$ for $n < N$.

For Chebyshev or Legendre polynomials the matrix Q is related to the simple matrix

$$\begin{bmatrix} 0 & 0 & 0 & 0 & & \cdots & & 0 \\ 1 & 0 & -1 & 0 & & & & \\ 0 & 1 & 0 & -1 & & & & \\ & & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & & \\ & & & & 1 & 0 & -1 & 0 \\ & & & & 0 & 1 & 0 & 0 \\ 0 & & \cdots & & 0 & 0 & 1 & 0 \end{bmatrix} \tag{25}$$

by row or column scaling, respectively. This matrix also corresponds to another Q obtained from polynomial basis functions which satisfy the recurrence relation $p_{n+1}(x) - p_{n-1}(x) = \int p_n(x)dx$.

Let λ be a nonzero eigenvalue of this matrix. The first N components $v_0, \ldots, v_{N-1}$ of the corresponding eigenvector may be then written as

$$v_n = i^{-n} \sin(n\phi) \tag{26}$$

with the last component being $v_N = -\cos(N\phi)\tan(\phi) i^N/2$. This follows from the formula

$$i \sin((n-1)\phi) - \frac{1}{i} \sin((n+1)\phi) = 2i \cos(\phi) \sin(n\phi) \tag{27}$$

and the requirement that $\sin(N\phi) = 0$. Nontrivial solutions can be obtained for $\phi = m\pi/N$ where $m = 1, \ldots, N-1$ (except that $m \neq N/2$). The corresponding eigenvalues $\lambda = 2i\cos(\phi)$ all lie on the imaginary axis. The remaining 2 (for odd N) or 3 (for even N) eigenvalues are zero, with the eigenvector p_N. Therefore, all eigenvalues lie on the imaginary axis and $|\lambda \leq 2$.

For Q obtained from the Chebyshev polynomial basis, we write the eigenvector components $v_0, \ldots, v_{N-1}$ as $v_n = i^{-n} Z_n$ and require $Z_0 = Z_N = 0$. The last component is then $v_N = i^{1-N} Z_{N-1}/(2N\lambda)$. We obtain the recurrence relation

$$Z_{n-1} + Z_{n+1} = -2in\lambda Z_n \tag{28}$$

whose solutions for nonzero λ are linear combinations of Bessel functions [1]

$$Z_n = \alpha J_n(\omega) + \beta Y_n(\omega) \tag{29}$$

where we have defined $\omega = -i/\lambda$.

We can obtain eigenvectors v provided that λ is a root of the equation

$$\det \begin{bmatrix} J_0(\omega) & Y_0(\omega) \\ J_N(\omega) & Y_N(\omega) \end{bmatrix} = 0 \tag{30}$$

or equivalently

$$\arg \left(\frac{H_0^{(1)}(\omega)}{H_N^{(1)}(\omega)} \right) = 0, \pi \tag{31}$$

where Hankel functions $H_n^{(1)}(\omega) = J_n(\omega) + iY_n(\omega)$ are introduced.

Remembering that $\arg(z) = \mathrm{im}(\log(z))$, we consider the imaginary part of the function

$$\Theta(\omega) = \frac{d}{d\omega} \log \left(\frac{H_0^{(1)}(\omega)}{H_N^{(1)}(\omega)} \right) \tag{32}$$

$$= \frac{H_1^{(1)}(\omega)}{H_0^{(1)}(\omega)} - \frac{H_{N-1}^{(1)}(\omega) - H_{N+1}^{(1)}(\omega)}{2H_N^{(1)}(\omega)} \tag{33}$$

which is approximately 1 over an interval of width $O(N)$ and decays towards zero for $\omega \gg N$. The plot of $\mathrm{im}(\Theta(\omega))$ for the case $N = 32$ is shown in figure 1. We conclude that in Chebyshev representation the eigenvalues $\lambda = -i/\omega$ are nearly evenly spaced in ω over a broad range.

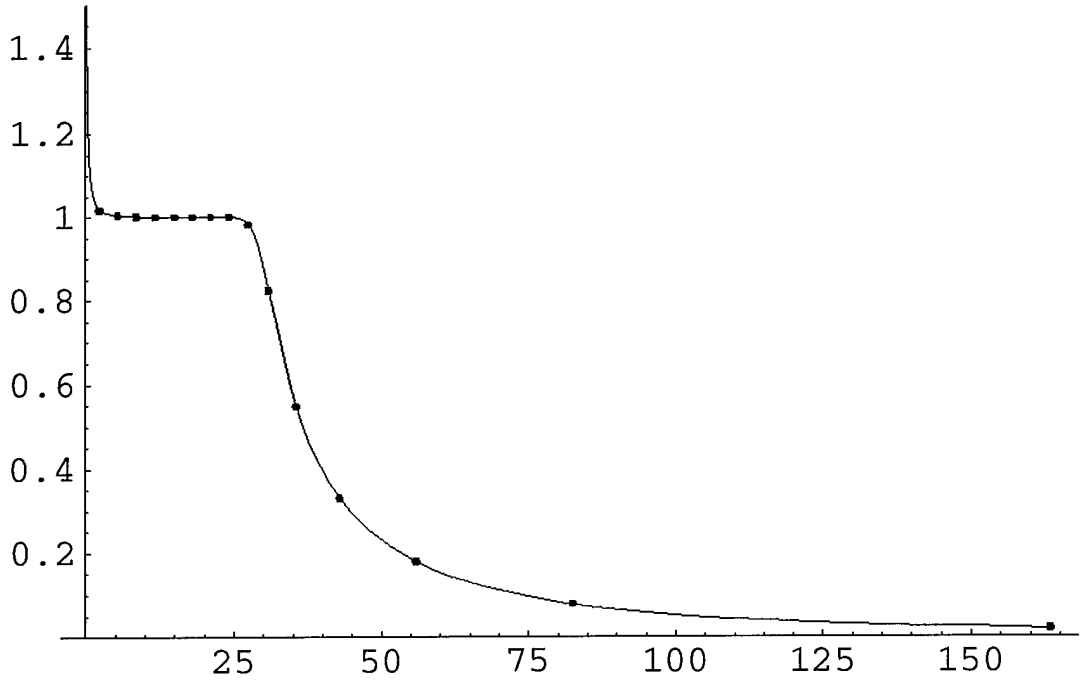

Figure 1: Derivative of the phase difference of Hankel functions vs. ω for the case $N = 32$. Dots are placed on the computed curve at ω values corresponding to the numerically computed eigenvalues $\lambda = -i/\omega$. The area under the curve between any two consecutive dots is π.

The smallest root ω corresponds to the largest λ. For $\omega \ll N$, the equation (31) reduces to $\arg(H_0^{(1)}(\omega)) \approx \pi/2$ and we conclude that $\min(|\omega|) \approx j_{0,1}$, the first zero of the Bessel function $J_0(\omega)$. The corresponding $\max(|\lambda|) \approx 1/j_{0,1} = 0.415831$.

Since $Qp_N = 0$, $\lambda = 0$ is an eigenvalue with multiplicity at least one. The basis of the invariant subspace associated with $\lambda = 0$ depends on the parity of N. For odd N, $\lambda = 0$ has multiplicity 2 and Q acts by mapping

$$N\,(1,0,2,0,2,0,\ldots,2,0)^T \stackrel{Q}{\longmapsto} p_N \stackrel{Q}{\longmapsto} 0 \quad , \tag{34}$$

while for even N, the zero eigenvalue has multiplicity three and one obtains the sequence

$$N\left(N^2/2,0,N^2-2^2,0,N^2-4^2,0,\ldots,N^2-(N-2)^2,0,0\right)^T \stackrel{Q}{\longmapsto}$$
$$N\,(0,2,0,2,0,\ldots,2,0)^T \stackrel{Q}{\longmapsto} p_N \stackrel{Q}{\longmapsto} 0. \tag{35}$$

This longer sequence is also responsible for a substantial increase in the condition number of the similarity transform S between the Q in Chebyshev representation and its Jordan decomposition. For example, $\kappa(S)$ is 36.3848 for $N = 63$, increases to 1.38497×10^6 for $N = 64$, and goes back down to 23.3846 for $N = 65$. In general, odd N produced better conditioned Jordan decomposition in all of the numerical tests, with $\kappa(S) \approx O(N/2)$. We shall pay particular attention to odd N of the form $2^J - 1$.

The physical interpretation of these sequences generated by the nilpotent part of Q for large N is that polynomial representations of functions whose first (and possibly second) integrals have discontinuities at the boundaries ($x = \pm 1$) are mapped to p_N, which is mapped to zero.

While this completes the description of the effect of Q in Chebyshev representation, the equation (31) could not be solved in closed form for all roots ω. Instead, the characteristic polynomial of Q was derived analytically in Mathematica 3.0 for $N = 32$, confirming that machine precision intervals around the numerically computed eigenvalues bracket the true roots. Numerical evidence suggests that all eigenvalues of Q obtained from the Chebyshev polynomial basis lie on the imaginary axis and that $|\lambda| \leq 0.415831$. This inequality is in excellent agreement with the analytic approximation (within machine precision for $N > 12$). Similarly, eigenvalues of Q arising from the Legendre polynomial basis are also on the imaginary axis, with $|\lambda| \leq 0.318310$.

Finally, the square of the Frobenius norm of Q in the limit $N \to \infty$ is given by

$$1 + \frac{1}{4} + \frac{1}{4(N-1)^2} + \frac{1}{4N^2} + \sum_{n=2}^{N-2} \frac{1}{2n^2} \longrightarrow \frac{\pi^2}{12} + \frac{3}{4} = 1.57247 \tag{36}$$

for the Q derived in the Chebyshev representation and by

$$1 + \frac{1}{3^2} + \sum_{n=2}^{N-1} \frac{2}{(2n+1)^2} \longrightarrow \frac{\pi^2}{4} - \frac{10}{9} = 1.35629 \tag{37}$$

for the Q derived in the Legendre representation.

6 Numerical results

Since Q is singular and its eigenvalues λ lie on the imaginary axis, for all real k the condition number of $I - kQ$ is

$$\kappa(I - kQ) = \sqrt{1 + (k \max(|\lambda|))^2} \tag{38}$$

which approaches $|k| \max(|\lambda|)$ for large $|k|$. The eigenvalues of $(I - kQ)^{-1}Q$ are of the form $\mu = \lambda/(1 - k\lambda)$. For real k one obtains $|\mu|^2 = |\lambda|^2/(1 + k^2|\lambda|^2)$. We conclude that $\max(\mu|) = \max(|\lambda|)$. Both Chebyshev and Legendre polynomial bases have small upper bounds on $|\lambda|$, so that the spectral radius of the proposed numerical scheme is well behaved for all real k.

The accuracy of the proposed method can be analyzed as follows. Let f_t represent the true solution satisfying the boundary condition and let $(D - kI)f_t = g$. When $k = 0$, the truncation error in representing $f_t(x)$ corresponds to setting the last coefficient g_N to zero. Similarly, the truncation error in representing the exact homogeneous solution $f_h(x)$ is also involved when $k \neq 0$. The error in determining f_p needs to be analyzed next.

Since $QD = I - \Pi_{P_0}$ and $DQ = I - \Pi_{P_{N-1}^\perp}$, writing $(I - kQ)f_p = Qg = (QD - kQ)f$ leads to

$$f_p = (I \quad kQ)^{-1}(I \quad kQ \quad \Pi_{P_0})f_t \tag{39}$$

so that

$$f_t - f_p = (I - kQ)^{-1}\Pi_{P_0}f_t \tag{40}$$

and since $D - kI = D(I - kQ) - \Pi_{P_{N-1}^\perp}$

$$(D - kI)f_p = g + k\Pi_{P_{N-1}^\perp}(I - kQ)^{-1}\Pi_{P_0}f_t. \tag{41}$$

The last equation follows from the observation that $D\Pi_{P_0} = 0$. Let $f_0 = (p_0, \Pi_{P_0}f_t)$ be the constant function component of f_t. Therefore, f_p is the exact solution of a modified problem where a term proportional to $p_N(x)f_0$ has been added to the right hand side.

When $k = 0$ or $f_0 = 0$, the method is exact apart from the truncation error. When $k \neq 0$ and $f_0 \neq 0$, the additional term in (41) follows from the expansion by minors, which is easily evaluated since Q is tridiagonal in representations of interest. For example, in Chebyshev representation, we obtain

$$2k\frac{(k/2)^N}{N!}c(k, N)p_N(x)f_0 \stackrel{\text{def}}{=} \varepsilon(k, N)p_N(x)f_0 \tag{42}$$

where

$$c(k, N) = \frac{1}{\det(I - kQ)} = \frac{1}{(-k)^{N+1}\chi_N(1/k)} \tag{43}$$

and $\chi_N(\cdot)$ is the characteristic polynomial of Q. The proportionality constant is given by

$$\varepsilon(k, N) = \frac{4}{(-2)^{N+1}N!\,\chi_N(1/k)} \tag{44}$$

and decreases rapidly as $N \to \infty$.

Let us consider the continuous analogue

$$h(x) - k\int h(x)dx = 1 \text{ where } \int h(x)dx \perp P_0 \tag{45}$$

and the exact solution

$$h(x) = \frac{e^{kx}}{I_0(k)}. \tag{46}$$

The Chebyshev polynomial representation of $h(x)$ is

$$h(x) = 1 + \sum_{n=1}^{\infty} \frac{2I_n(k)}{I_0(k)} T_n(x) \tag{47}$$

where $I_\nu(k)$ are the modified Bessel functions. For large enough N the truncation error becomes small and $h_{N+1}| < |h_N| \leq \epsilon$. Therefore, whenever $|k| \leq N$

$$(I - kQ)h = 1 - k\frac{h_N p_N}{2(N-1)} - k\frac{h_{N+1}p_{N+1}}{2N} \approx 1 \tag{48}$$

so that h is approximately the solution of the discrete problem as well. We obtain

$$\varepsilon(k, N) \approx \frac{2kI_N(k)}{I_0(k)} \text{ when } |k| \leq N. \tag{49}$$

However, the range of applicability of this approximation is too narrow since the term $\varepsilon(k, N)$ may be small even if $|k| \gg N$. The resolution criterion $|\varepsilon(k, N)| \leq \epsilon$ leads to

$$|\chi_N(1/k)| \geq \frac{4}{\epsilon N!\, 2^{N+1}} \approx \frac{4}{\epsilon}\left(\frac{e}{2(N+1)}\right)^{N+1}\sqrt{\frac{N+1}{2\pi}} \tag{50}$$

where the approximation follows from Stirling's asymptotic formula for $N! = \Gamma(N+1)$ applicable for $N \gg 1$. This criterion can be simplified as follows.

Clearly $c(0, N) = 1$. For each N, nonzero eigenvalues of Q occur in conjugate pairs on the imaginary axis and $\det(I - kQ)$ is the product of the terms of the form $1 + k^2|\lambda_j|^2 > 1$, showing that $|c(k, N)| \to 0$ as $|k \to \infty$. One would expect that the solution k to $|\varepsilon(k, N)| \leq \epsilon$ depends on the ability to resolve the thin boundary layer near x_b. Indeed, by curve fitting the numerical results for $N = 4, 5, \ldots, 128$ one obtains the criterion that machine precision $\epsilon = 2.22045 \times 10^{-16}$ is approximately reached provided that $N \geq 2.06131 + 8.53338\sqrt{1 + |k|}$.

Truncation error in representing $f_h(x)$ must also be considered. The resolution requirements are virtually identical. For Chebyshev polynomial basis and $|k| \gg 1$, $f_h(x)$ can be represented to machine precision provided that the last coefficient of f_h (which is given in terms of the modified Bessel function $I_N(\cdot)$) satisfies $2I_N(|k|)e^{-|k|} \leq \epsilon$. This leads to the approximate criterion $N \geq 2.81178 + 8.05974\sqrt{1 + |k|}$, according to asymptotic analysis and curve fitting to numerical experiments.

The simplified combined criterion

$$N \geq 3 + 9\sqrt{|k| + 1} \Leftrightarrow k| \leq \frac{(N-3)^2}{81} - 1 \text{ and } N \geq 12 \tag{51}$$

works well in practice (figure 2). The criterion (51) has been verified for $12 \leq N \leq 256$, where $|\varepsilon(k,N)| \leq 6.60738 \times 10^{-15}$ and the truncation error is even lower. For $N > 256$, the resolution requirements are only slightly more stringent, since the worst case residual error appears to be slowly growing with N. A more complicated criterion of the form

$$\sqrt{|k|+1} \leq \frac{N}{10} + \frac{4}{5}\log(N) - 3 \tag{52}$$

fits the numerical results almost exactly for $128 \leq N \leq 256$ and may be extrapolated to larger N. Our numerical results indicate that the criterion (52) produces worst case residues $|\varepsilon(k,N)|$ of less that 100ϵ at $N = 4096$ and less than 10^{-10} at $N = 2^{16} = 65536$.

Numerical tests in Chebyshev polynomial representation were done at $N = 16, 32, 64$ and $k = -10, -1, 0, 1, 10$ and compared to the analytically obtained integrals for $g(x) = T_n(x)$ where $n = 1, 2, 4, 8, \ldots, N/2$. Standard precision was used to obtain the polynomial approximation of the solution but the exact integrals

$$f(x) = \int_{x_b}^{x} e^{k(x-s)} g(s) ds \tag{53}$$

were obtained analytically and then evaluated in extended precision using Mathematica 3.0, which was essential to capture the delicate cancellations in the exact solutions. For example, the exact solution of

$$\frac{df(x)}{dx} - f(x) = T_{32}(x) \tag{54}$$

with the boundary condition $f(1) = 0$ satisfies

$$\begin{aligned} f(0) \;=\; & 1523681485112275626378695517688630699203508225/e - \\ & 560531093266377270849883382873867646780763137 \end{aligned} \tag{55}$$

which is about -0.00062, i.e. 48 orders of magnitude less than the terms involved.

In 63 of the 75 tests, the L_2 and L_∞ relative norms of the error at the Chebyshev collocation points were below 10^{-14}. The remaining 12 results are summarized in Table 1. These results match the above analysis. For example, when $|k| = 10$ and $N = 16$, the relative truncation error in representing $e^{k(x-x_b)}$ is 2.73×10^{-6} and does not become negligible until $N = 32$.

7 Generalizations

The method presented here may be generalized to higher order equations in one variable of the form

$$\sum_{j=0}^{J} D^j A_j f = g \tag{56}$$

rewritten as

$$\sum_{j=0}^{J} Q^{J-j} A_j f_p = Q^J g = \sum_{j=0}^{J} Q^J D^j A_j f \tag{57}$$

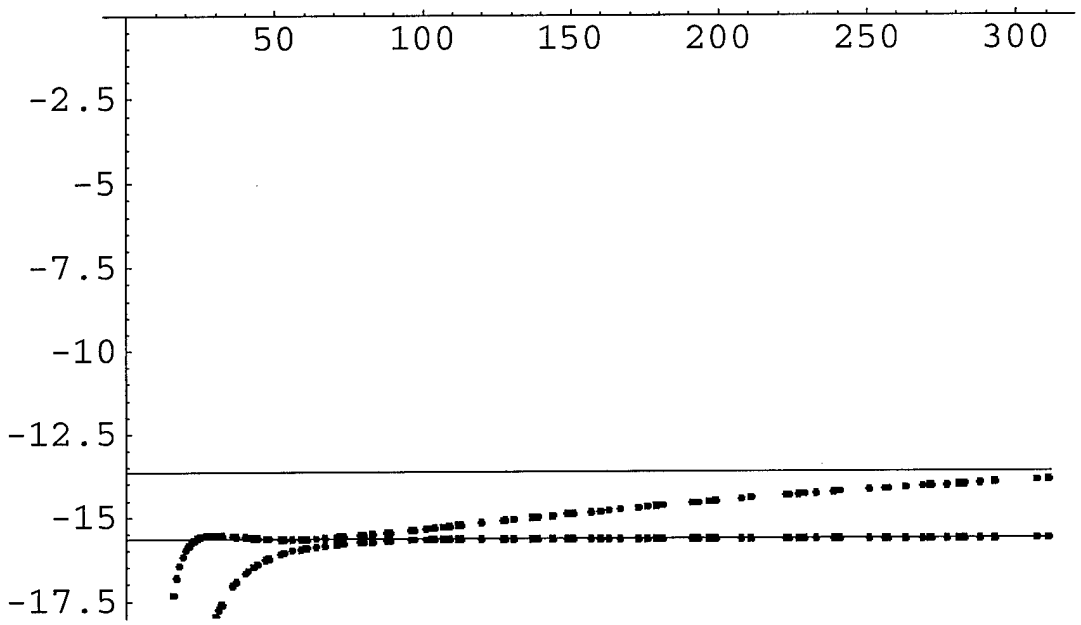

Figure 2: The maximum residual $\log_{10}|\varepsilon(k(N), N)|$ vs. N for the worst case $k(N)$ according to the resolution criteria (51) [upper points] and (52) [lower points]. The criterion (51) gives residues closer to ϵ for $N \leq 80$ and remains usable at somewhat larger N, but the slightly more conservative criterion (52) is closer to achieving full precision for $N > 80$. Horizontal lines indicate the machine precision $\epsilon = 10^{-15.6536}$ and the error tolerance set at 100ϵ.

N	k	n	$\|\delta\|_2/\|f\|_2$	$\|\delta\|_\infty/\|f\|_\infty$
16	-10	1	8.55882×10^{-7}	8.71212×10^{-7}
16	-10	2	1.28628×10^{-6}	2.01382×10^{-6}
16	-10	4	3.03513×10^{6}	3.05528×10^{6}
16	-10	8	7.85599×10^{-5}	8.82017×10^{-5}
16	1	8	4.96642×10^{-13}	4.46525×10^{-13}
16	1	8	4.97129×10^{-13}	4.47325×10^{-13}
16	10	1	8.55882×10^{-7}	8.71212×10^{-7}
16	10	2	1.28628×10^{-6}	2.01382×10^{-6}
16	10	4	3.03513×10^{-6}	3.05528×10^{-6}
16	10	8	7.85599×10^{-5}	8.82017×10^{-5}
32	-10	16	1.11479×10^{-12}	1.09573×10^{-12}
32	10	16	1.11461×10^{-12}	1.09789×10^{-12}
other 63 tests			$< 10^{-14}$	$< 10^{-14}$

Table 1: Relative error in the L_2 and the L_∞ norm at the Chebyshev collocation points for each resolution.

so that f_p becomes

$$f_p = \left(\sum_{j=0}^{J} Q^{J-j} A_j \right)^{-1} Q^J g. \tag{58}$$

Invertibility depends on the specific A_j. Furthermore, J boundary conditions are to be satisfied by adding a linear combination of J precomputed homogeneous solutions. Alternatively, problems of the form

$$\prod_{j=1}^{J} (D - A_j) f = g \tag{59}$$

can be converted into a sequence of first order problems

$$f_{pj} = (I - Q A_j)^{-1} Q g_j \tag{60}$$

where $g_1 = g$, $f_J = f$ and $g_j = f_{j-1}$ for $j = 2, \ldots, J$. In addition to invertibility requirements, this approach requires precomputing the homogeneous solutions of the J subproblems. Boundary conditions on each $f_j = f_{pj} + \alpha_j f_{hj}$ must also be available.

In multidimensional domains, the number of homogeneous solutions is proportional to the number of boundary points. Therefore, the proposed method is most beneficial in low dimensional problems where the differential operators have a favorable structure. However, some multidimensional problems may be factored into a sequence of one dimensional problems, for which the proposed method can be very effective. In fact, this investigation was motivated by a factorization of the 2D Laplacian. This example shows that non-normal operators of type (9) arise naturally in solving partial differential equations as well.

8 Conclusion

The advantages of the proposed method include numerical stability, high accuracy and efficiency. The pseudo-inverse of the derivative operator has a simple tridiagonal form in commonly used polynomial representations and $I - kQ$ is easily inverted for real k. The boundary condition is applied afterwards. By contrast, the Lanczos tau method applies the boundary condition by introducing a full row into the matrix to be inverted, and thus increases the computational effort by a factor of $O(N)$.

Our analysis and numerical experiments indicate that the pseudo-inverse of D and delayed application of boundary conditions are most useful in one-dimensional problems and low multidimensional problems which may be factored into one-dimensional subproblems.

References

[1] Carl M. Bender and Steven A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill Book Company, New York, 1978.

[2] E. A. Coutsias, T. Hagstrom, J. S. Hesthaven and D. Torres, *Integration preconditioners for differential operators in spectral τ-methods*, Proc. of International Conference on Spectral and High Order Methods, Houston, USA, 1995.

[3] Gene H. Golub and Charles F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, Maryland, 1983.

[4] David Gottlieb and Steven A. Orszag, *Numerical Analysis of Spectral Methods*, SIAM, Philadelphia, 1977.

[5] L. N. Trefethen, *Pseudospectra of matrices*, in D. F. Griffiths and G. A. Watson, Numerical Analysis 1991, Longman Sci. Tech. Publ., 1992, pp. 234-266.

[6] S. V. Tsynkov, *An Application of Nonlocal External Conditions to Viscous Flow Computations*, J. Comput. Phys., 116 (1995), pp. 212-225.